

DOCUMENT INFORMATION CLASSIFYING METHOD, AND METHOD AND SYSTEM FOR DOCUMENT INFORMATION COLLECTION USING THE SAME

Publication number: JP7049875

Publication date: 1995-02-21

Inventor: YUASA HIROKO; KOJIMA KEIJI

Applicant: HITACHI LTD

Classification:

- International: G06F12/00; G06F17/21; G06F17/27; G06F17/30; G06F12/00; G06F17/21; G06F17/27; G06F17/30; (IPC1-7): G06F17/30; G06F12/00; G06F17/27

- European:

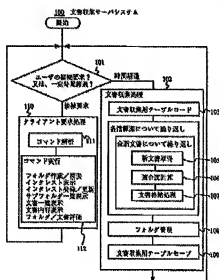
Application number: JP19930195839 19930806

Priority number(s): JP19930195839 19930806

Report a data error here

Abstract of JP7049875

PURPOSE: To determine a corresponding folder and obtain proper information classifications by classifying information in consideration of the degree of matching between retrieval conditions which are stored corresponding to respective folders and the hierarchical structure of the retrieval conditions. **CONSTITUTION:** A document collecting server system 100 accesses external information sources periodically and starts a document collecting process 102. When a table for document collecting is loaded, documents stored in the respective information sources are acquired and documents meeting the retrieval condition group that the user has registered are retrieved. At this time, the frequency of appearance of each word in the retrieval conditions in an object document is decided as the degree of matching between the object document and retrieval conditions through matching degree calculation 106. Then a document storing process 107 selects a folder where the object document is classified in consideration of the hierarchical structure of the folders among folders where the matching retrieval conditions are registered and stores it inside. Further, a folder managing process 108 rearranges documents corresponding to the gathering state of documents by automatically dividing the folder in which many documents are stored.



特開平7-49875

(43) 公開日 平成7年(1995)2月21日

(51) Int.Cl. ⁴ G 0 6 F 17/30 12/00 17/27	識別記号 5 2 0 A 8944-5B	庁内整理番号 9194-5L 7315-5L	F I	技術表示箇所 G 0 6 F 15/ 401 15/ 20 3 1 0 D 5 5 0 F
審査請求 未請求 請求項の数28 O L (全 19 頁)				

(21) 出願番号 特願平5-195839

(22) 出願日 平成5年(1993)8月6日

(71) 出願人 00005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 湯浅 寛子

東京都国分寺市東意ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 小島 啓二

東京都国分寺市東意ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

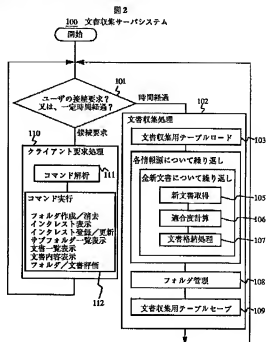
(74) 代理人 弁理士 小川 勝男

(54) 【発明の名称】 文書情報分類方法およびそれを用いた文書情報収集方法、文書情報収集システム

(57) 【要約】 (修正有)

【構成】 文書収集サーバシステム100は、自動的に複数の情報源に接続して新文書を取得し、適合度計算106によって、あらかじめユーザが記述した検索条件との適合度を調べる。文書格納処理107は、検索条件間の関係から分類体系を構成し、適合した文書を分類してフォルダに格納する。フォルダ管理処理108は、各フォルダへの情報の集まり具合を監視し、自動的にフォルダの細分化、統合、構造の変更を行なって情報の整理をする。

【効果】 各分類への情報の集まり具合に応じて、分類体系や検索条件を改善し、各分類に分類される情報量をその全体を容易に把握できる程度の数に抑えることができる。



1

【特許請求の範囲】

【請求項1】 計算機により情報を自動的に分類するための方法であって、

階層関係で相互に関連付けられた複数のフォルダの一つにそれぞれ対応し、それぞれ検索すべき一つまたは複数の単語を指定する複数の検索条件を記憶し、

各フォルダに対応して記憶された検索条件と分類すべき文書情報との間の適合度を検出し、

各フォルダと該情報との間の検出された適合度と上記階層関係とに基づいて、該複数のフォルダの内、該文書情報を登録すべき一つまたは複数のフォルダを決定し、該決定された一つまたは複数のフォルダに対して該文書情報を記憶する情報分類方法。

【請求項2】 該検出は、

該文書情報と、各フォルダに対応して記憶された検索条件が指定する単語の各々との間の適合度を検出し、

各フォルダに対応して記憶された検索条件が指定する単語のそれぞれと該文書情報との間に関して検出された適合度の総和を、そのフォルダと該文書情報との間の適合度として決定するステップを含む請求項1記載の文書情報分類方法。

【請求項3】 該文書情報と各フォルダに対応して記憶された検索条件に含まれる一つまたは複数の単語の各々との間の適合度を決定するステップは、その文書情報内でのその単語の重みと、その検索条件内でのその単語の重みとの積をその単語とその文書情報との適合度として決定するステップを有する請求項2記載の文書情報分類方法。

【請求項4】 各フォルダに対応して記憶された検索条件に含まれる一つまたは複数の単語の各々のその文書情報内での重みは、その文書情報内でのその単語の出現回数に比例する値である請求項3記載の文書情報分類方法。

【請求項5】 各フォルダに対応して記憶された検索条件に含まれる一つまたは複数の単語の各々のその検索条件内の重みは、その検索条件内の複数の単語内でのその単語の出現回数に比例する値である請求項3または4記載の文書情報分類方法。

【請求項6】 該決定は、

該文書情報と各フォルダとの間に対して検出された適合度により、該文書情報に適合する一つまたは複数のフォルダを検出し、

該検出により複数のフォルダが検出されたときには、それらの検出された複数のフォルダ内の該文書情報を対応させる一つまたは複数のフォルダを、該検出された複数のフォルダの間、上記階層関係内での相対的位置関係に依って選択する請求項1記載の文書情報分類方法。

【請求項7】 該選択は、該文書情報に適合すると検出された該複数のフォルダの中に、相対的に上下関係にある一群のフォルダが含まれているときには、該一群のフォルダを代表する一つのフォルダを該文書情報に対応させ

2

るフォルダとして選択するステップを有する請求項6記載の文書情報分類方法。

【請求項8】 該選択は、該検出された複数のフォルダの中に、該一群のフォルダと相対的に上下関係にはない他の一群のフォルダが含まれているときには、該他の一群のフォルダを代表する一つのフォルダを該文書情報に対応させる他のフォルダとして検出するステップをさらに有する請求項7記載の文書情報分類方法。

【請求項9】 該一群のフォルダを代表する一つのフォルダは、該一群のフォルダの内、最下層に位置するフォルダである請求項7記載の文書情報分類方法。

【請求項10】 計算機により文書情報を自動的に分類するための方法であって、

階層関係で相互に関連付けられた複数のフォルダの一つにそれぞれ対応し、それぞれ一つまたは複数の検索すべき単語を指定する複数の検索条件を記憶し、

各フォルダに対応して記憶された検索条件と予め決定した判断基準とに基づいて、分類すべき文書情報を対応させるフォルダとして、該複数のフォルダの一つまたは複数の

決定されたフォルダに対応して該文書情報を記憶し、複数の分類すべき文書情報の各々に対して上記決定および記憶を行ない、

各フォルダに対応して記憶された複数の文書情報が、そのフォルダの再構成のために定めた所定の条件を満たすか否かを判別し、

いずれか一つのフォルダが該所定の条件を満たしたとき、その一つのフォルダに対応して記憶された複数の文書情報とそのフォルダに対応して記憶された検索条件を再構成するステップを有する文書情報分類方法。

【請求項11】 該所定の条件は、該一つのフォルダの登録された文書情報の総数が所定値を越えているということである請求項10記載の文書情報分類方法。

【請求項12】 該再構成するステップは、

そのフォルダに対応して記憶された検索条件を、その検索条件が指定する複数の単語の一部をそれぞれ指定する複数の新たな検索条件に分割し、

該一つのフォルダに登録された複数の文書情報を複数の文書情報群に分割し、

該一つのフォルダを新たな複数のフォルダで置換し、該複数の新たなフォルダの各々に対応して、該新たな複数の検索条件の一つで指定される一つまたは複数の単語と、該複数の文書情報の分割で得られた一部の文書情報を記憶するステップを有する請求項10記載の文書情報分類方法。

【請求項13】 該複数の文書情報を分割するステップは、

該一つのフォルダに対応して記憶された検索情報を分割して得られた複数の新たな検索条件に適合する文書情報からなる複数の文書情報部分群に分割するステップから

3

なる請求項1記載の文書情報分類方法。

【請求項14】該再構成するステップは、

該一つのフォルダに登録された複数の文書情報の一部とそのフォルダに対応して記憶された検索条件が指定する複数の単語の一部とを選択し、

該一のフォルダの下位の階層に、少なくとも一つの新たなフォルダを配置し、

該新たなフォルダに対応して、該選択された一部の単語と該選択された一部の文書情報を記憶するステップを有する請求項10記載の文書情報分類方法。

【請求項15】該複数の文書情報の一部と該一部の単語を選択するステップは、

該一つのフォルダに対応して該複数の文書情報を、該一つのフォルダに対応して記憶された単語群の一部で検索可能であるが、該単語群のうちの他の一部の単語では検索不可能な一部の文書情報と、該一部の単語および該他の一部の単語のいずれでも検索可能な他の文書情報とに分離し、

該分離で得られた一部の文書情報およびその分離に使用した該一部の単語を選択するステップを有する請求項14記載の文書情報分類方法。

【請求項16】該再構成するステップは、

新たなフォルダを生成し、

該一つのフォルダと他の一のフォルダに重複して登録された複数の文書情報を該新たなフォルダに対応して登録し、該一のフォルダと該他の一のフォルダに対する、該重複する複数の文書の登録を削除し、

該一つのフォルダに対応して記憶された単語群と該他の一つのフォルダに対応して記憶された他の単語群の内の、該重複する複数の文書を検索するための単語群を、該新たなフォルダに対応して記憶するステップを有する請求項10記載の文書情報分類方法。

【請求項17】該新たなフォルダは、該一つのフォルダと該他の一のフォルダと同じ階層に配置される請求項16記載の文書情報分類方法。

【請求項18】該文書情報に対応させるための一つまたは複数のフォルダの決定は、

各フォルダに対応して記憶された検索条件が指定する単語群に基づいて、該分類すべき文書情報と各フォルダとの間の適合度を検出し、

各フォルダと該文書情報との間の検出された適合度と上記階層関係とに基づいて、該文書情報を登録するためのフォルダとして、該複数のフォルダの一つまたは複数を選択するステップを有する請求項10記載の文書情報分類方法。

【請求項19】計算機により文書情報を自動的に分類するための方法であって、

階層関係で相互に関連付けられた複数のフォルダの一つにそれぞれ対応し、それぞれ一つまたは複数の検索すべき単語を指定する複数の検索条件を記憶し、

4

各フォルダに対応して記憶された検索条件と予め定めて判断基準とに基づいて、分類すべき文書情報に対応させるフォルダとして、該複数のフォルダの一つまたは複数を選択し、

決定されたフォルダに対応して該文書情報を記憶し、複数の分類すべき文書情報の各々に対して上記決定および記憶を行ない、

該複数のフォルダの内の一部の複数のフォルダに対応して記憶された複数の文書情報が、該複数のフォルダの再構成のために定めた所定の条件を満たすか否かを判別し、

いずれかの一の複数のフォルダが該所定の条件を満たしたとき、該一部の複数のフォルダに対応して記憶された複数の文書情報と、該一部の複数のフォルダに対応して記憶された複数の検索条件を再構成するステップを有する文書情報分類方法。

【請求項20】該所定の条件は、該一部の複数のフォルダが、上位側のフォルダとそのフォルダの下位側のフォルダに関する条件を含み、

該再構成は、

該下位側のフォルダに対応して登録された文書情報群と該上位側のフォルダに対応して登録された文書情報群とを、該上位側のフォルダと該下位側のフォルダに対して分配し直し、

この分配し直しの後に、該下位側のフォルダに対応して登録された新たな文書情報群と該上位側のフォルダに登録された新たな文書情報群とに基づいて、該下位側のフォルダに対応して登録された文書情報群と該上位側のフォルダと該下位側のフォルダに対して分配し直すステップを有する請求項19記載の文書情報分類方法。

【請求項21】該所定の条件は、該下位側のフォルダに対応して登録された文書情報の数と該上位側のフォルダに対応して登録された文書情報の数の相対的大さに関する条件である請求項19記載の文書情報分類方法。

【請求項22】該条件は、該下位側のフォルダに対応して登録された文書情報の数が、該上位側のフォルダに対応して登録された文書情報の数より少ないことである請求項21記載の文書情報分類方法。

【請求項23】該文書情報に対応させるための一つまたは複数のフォルダの決定は、

各フォルダに対応して記憶された検索条件とに基づいて、分類すべき文書情報と各フォルダとの間の適合度を検出し、

各フォルダと該文書情報との間の検出された適合度と上記階層関係とに基づいて、該文書情報に対応させるフォルダとして、該複数のフォルダの一つまたは複数を選択するステップを有する請求項19記載の文書情報分類方法。

【請求項24】データベースを保持する記憶装置と、該計

5

算機からユーザが指定した文書情報を選択的に検索する計算機とを有する計算機システムにおいて、

階層関係で相互に関連付けられた、ユーザが指定した複数のフォルダの一つにそれぞれ対応し、それぞれ一つまたは複数の検索すべき単語を指定する複数の検索条件を記憶し、

そのデータベースに新規に登録される文書情報があるかを監視し、

新規に登録された文書情報があるときには、その文書情報と各検索条件との適合度を判別し、

各検索条件と該文書情報との適合度と該階層関係とに基づいて、該文書情報に対応させる一つまたは複数のフォルダを決定する文書情報収集方法。

【請求項25】該複数のフォルダの各々が、そのフォルダに対応して記憶された複数の文書情報に関連するフォルダの再構成に関する条件を満たすか否かを判別し、

いずれか一つのフォルダが該再構成の条件を満たすとき、新たなフォルダを生成し、

その一つのフォルダに対応して記憶された複数の文書情報の少なくとも一部を検索するための新たな検索条件を生成し、

該新たな検索条件と該一部の文書情報を該新たなフォルダに対応して記憶するステップをさらに有する請求項24記載の文書情報収集方法。

【請求項26】いずれかのフォルダが、そのフォルダに対応して記憶された複数の文書情報を複数の新たなフォルダに分割するための分割条件を満たすか否かを判別し、

いずれか一つのフォルダが該分割条件を満たすとき、複数の新たなフォルダを生成し、

その一つのフォルダに対応して記憶された複数の文書情報および検索条件とから、該複数の文書情報を複数群に分割するための複数の検索条件を決定し、

該分割により得られた複数群の文書情報の内の一つの群の文書情報と、該分割により得られた複数群の検索文書情報の内の一つの群の文書情報とを、該新たな複数のフォルダの一つに対応して記憶するステップをさらに有する請求項24記載の文書情報収集方法。

【請求項27】ある一組の複数のフォルダが、それらに重複して対応して記憶された複数の文書情報を分離して記憶するための条件を満たすか否かを判別し、

その一組のフォルダが、該分離条件を満たすとき、新たなフォルダを生成し、

それらのフォルダに重複して記憶された複数の文書情報を検索するための検索条件を、該一組のフォルダのいずれかにそれぞれ対応して記憶された検索条件に基づいて生成し、

該新たなフォルダに対応して、該生成された検索条件を記憶し、

該重複する複数の文書情報を、該新たなフォルダに対応

6

するように記憶し直すステップをさらに有する請求項24記載の文書情報収集方法。

【請求項28】ユーザに提供すべき文書情報を含むデータベースを記憶する手段を有する第1の計算機と、

該第1の計算機と交信して、該データベース内の文書情報を検索するための第2の計算機と、

該第2の計算機に接続された、ユーザが操作可能な端末とを有し、

該端末は、相互に階層関係で関係づけられた、ユーザが指定した複数のフォルダの名称と、それぞれのフォルダに対応してユーザが指定した複数の単語を含む複数の検索条件とを該第2の計算機に送付する手段を有し、

該第2の計算機は、該送付された複数のフォルダの名称と検索条件を記憶する手段と、

そのデータベースに新規に登録される文書情報があるかを該第1の計算機と交信して検出する手段と、

新規に登録された文書情報があるときには、その文書情報と該複数の検索条件との適合度を判別し、各検索条件と該文書情報との適合度と該階層関係とに基づいて、該文書情報に対応させる一つまたは複数のフォルダを決定し、決定されたフォルダの各々に対応して該文書情報とその名称を記憶する手段と、

該端末からの要求に応じて、該複数のフォルダの名称とそれぞれのフォルダに対応して記憶された複数の文書情報の名称とを該端末に送付する手段とを有し、

該端末は、該送付された複数のフォルダの名称を有する複数のフォルダを、該階層関係が識別可能な態様で表示し、該送付された複数の文書情報の名称を、それぞれの文書情報に対応するフォルダに対応して表示する手段をさらに有する文書情報収集システム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、計算機ネットワークを介して、自動的に情報を収集、分類、整理する情報収集システムに関する。

【0002】

【従来の技術】計算機ネットワークの整備は急速に進んでおり、オンライン情報検索サービス、ネットニュースからの情報収集、電子メールや電子掲示板を利用した質疑応答といった、いわゆる情報のブロードキャッチが行なえる環境が整いつつある。

【0003】これらの最新情報の有用性は認識されているものの、次のような点が問題となり、有効に利用されていない。

【0004】(1) 情報源によって利用法が異なり、複数の情報源から情報収集する操作が煩雑である。

【0005】(2) 検索式を論理式で入力しなければならぬ。所望の情報を得るための適切な検索式を記述するのは難しい。

【0006】(3) 収集した情報の分類と整理に手間と時間がかかる。

【0007】「21世紀の情報化社会」(日経バイト1991年11月320ページ〜331ページ)に記載されている広域検索サーバWAISは、(1)の問題点をプロトコルを共通化(NISO 739.50を拡張)し、さらに情報源への接続と検索を自動化することにより解決し、

(2)の問題点を関連性フィードバックにより解決した。関連性フィードバックは次のような検索条件の精練手法である。ユーザが検索したい内容を記述すると、それを検索条件としてWAISはその内容に合う情報を検索し提示する。ユーザがその中から欲しかった情報を選ぶと、WAISはユーザが選んだ情報を検索条件にフィードバックし、検索条件を改善する。この関連性フィードバックを用いた情報検索により、ユーザは検索式を記述することなく所望の情報を検索できるようになった。

【0008】(3)の問題点を解決するために、様々な文書の自動分類システムが考案されている。

【0009】たとえば、特開平1-188934の文書分類システムは、標本文書群を調べることにより、各分野におけるキーワードの出現頻度情報を得て、入力された文書からキーワードを抽出して、分野毎に点数を計算し、最高得点の分野へ分類する。

【0010】特開昭63-214832の通知文書処理システムは、通知文書の書式を解析し、通信文中に出現する単語の重みを分類カテゴリ別に付加し、その総和を求め、最大となるカテゴリを選ぶことにより分類する。

【0011】

【発明が解決しようとする課題】WAISは、上記(1)、(2)の問題点を解決したが、収集した情報の分類、整理に関しては配慮していない。

【0012】階層的に情報を分類整理することが望まれるが、従来の方法では、これに適していなかった。

【0013】また、(3)を解決する従来の自動分類システムにおいては、分類する分類体系をあらかじめ確立しておく必要があった。さらに、各分野を特徴付けるキーワード群やキーワード群の出現頻度などをあらかじめ与えるか、または求めるかする必要があった。

【0014】しかし、あらかじめ適切な汎用的分類体系を設けるのは困難である。分類体系が適切でないと、ある分類に多くの情報が集中することがある。ある分類の情報が多くなり過ぎると、ユーザは収集した情報の全容を把握しにくくなる。

【0015】また、最先端の分野では多くの人に認められる分類体系や専門用語が確定していないことが多く、しかも頻繁に変更される。最先端の分野に関する文書を従来の自動分類システムで適切に分類するのは難しい。

【0016】本発明の第1の目的は、階層的に情報を分類整理するのに適した文書情報分類方法、それを使用し

た文書情報収集方法およびシステムを提供することにある。

【0017】本発明の第2の目的は、収集した文書情報の集まり具合から、分類体系と分類に用いる検索条件の改良を自動的にこなす文書情報収集方法およびシステムを提供することにある。

【0018】

【課題を解決するための手段】本発明による第1の文書情報分類方法は、階層関係で相互に関連付けられた複数のフォルダの各々に対応して、一つまたは複数の検索条件からなる検索条件群を記憶し、各フォルダに対応して記憶された検索条件群に基づいて、分類すべき情報と各フォルダとの間の適合度を検出し、各フォルダと該情報との間の検出された適合度と上記階層関係とに基づいて、該情報が対応するフォルダとして、該複数のフォルダの一つまたは複数の数を決定し、該決定された一つのフォルダまたは複数のフォルダの各々に対応して該情報を記憶するステップを有する。

【0019】本発明による第2の文書情報分類方法は、階層関係で相互に関連付けられた複数のフォルダの各々に対応して、一つまたは複数の検索条件からなる検索条件群を記憶し、各フォルダに対応して記憶された検索条件群と予め定められた基準とに基づいて、分類すべき情報を対応させるフォルダとして、該複数のフォルダの一つまたは複数の数を決定し、決定されたフォルダに対応して該情報を記憶し、複数の分類すべき情報の各々に対して上記決定および記憶を行ない、各フォルダに対応して記憶された複数の情報が、そのフォルダの再構成のために定めた所定の条件を満たすか否かを判別し、いずれか一つのフォルダが該所定の条件を満たしたとき、その一つのフォルダに対応して記憶された複数の情報とそのフォルダに対応して記憶された一群の検索条件を再構成するステップを有する。

【0020】本発明による第3の文書情報分類方法は、階層関係で相互に関連付けられた複数のフォルダの各々に対応して、一つまたは複数の検索条件からなる検索条件群を記憶し、各フォルダに対応して記憶された検索条件群と予め定められた基準とに基づいて、分類すべき情報を対応させるためのフォルダとして、該複数のフォルダの一つまたは複数の数を決定し、決定されたフォルダに対応して該情報を記憶し、複数の分類すべき情報の各々に対して上記決定および記憶を行ない、該複数のフォルダの内の一部の複数のフォルダに対応して記憶された複数の情報が、該複数のフォルダの再構成のために定めた所定の条件を満たすか否かを判別し、いずれか一部の複数のフォルダが該所定の条件を満たしたとき、該一部の複数のフォルダに対応して記憶された複数の情報と、該一部の複数のフォルダに対応して記憶された一群の検索条件を再構成するステップを有する。

【0021】

9

【作用】本発明による第1の文書情報分類方法では、各フォルダに対応して記憶された検索条件と検索対象文書情報との適合度と、複数の検索条件の階層構造とを考慮して、検索対象文書情報に対応させるフォルダを決定するので、ユーザが記述した検索条件群を階層構造をなす分類体系であると見做して収集した文書情報を分類できる。

【0022】本発明による第2の文書情報分類方法では、各フォルダに対応して記憶された文書情報に依存して、各フォルダの分割など、フォルダの再構成をすること
10
が出来る。したがって、検索により得られた文書情報の集まり具合に応じて、自動的に分類体系を変更できる。

【0023】本発明による第3の文書情報分類方法では、複数のフォルダにまたがるフォルダの再構成をすることが出来る。

【0024】

【実施例】以下本発明の1実施例について説明する。

【0025】本実施例の文書情報収集システムが対象とするのは、オンライン文書情報検索サービス、電子メール、電子掲示板などを介して電子的に得ることができ、それぞれユーザにとって意味のある内容を一群の文
20
字で表した情報である。以下のような情報を文書情報とよぶ。

【0026】これらのサービスは、それぞれ様々な企業や団体により運営されている。以後これらのサービスを情報源と呼ぶ。各情報源が提供する文書情報は、一般に、多岐に亘るので、複数の分野に分けてユーザに提示される。これらの分野をドメインと呼ぶ。ドメインにおいて提供される個々の情報を文書と呼ぶ。文書が検索条件に適合したときに格納する検索結果格納領域をフォルダと呼ぶ。

【0027】図1に本実施例の文書収集システムと本実施例の文書収集システムが文書収集する外部の情報源とからなるシステム構成例を示す。本実施例の文書収集システムは文書収集クライアント500と文書収集サーバ510とからなる。

【0028】文書収集クライアント500はネットワーク上に複数存在して同時に文書収集サーバ510にアクセスすることが出来る。

【0029】文書収集クライアント500のメモリ522上の文書収集クライアントシステム501は、ユーザが、収集した文書を格納するフォルダを作成したり、どのような文書を収集するかを表す検索条件を各フォルダに登録したり、フォルダに収集された文書を見たりするためのグラフィカル・ユーザ・インタフェースを提供する。

【0030】文書収集サーバ510のメモリ523上の文書収集サーバシステム100は、文書収集クライアントシステム501からの要求に応じて文書情報を提供す
50

10

る一方で、自動的に、ニュースサーバ520や文書サーバ521などの外部の情報源から、ユーザが登録した検索条件群に適合する文書を収集し、さらに分類、整理を行う。

【0031】まず、文書収集クライアントシステム501について説明する。

【0032】ユーザが文書収集クライアントシステム501を起動すると図3に示すようなインタフェース画面400をCRT502上に表示する。ユーザはこのインタフェース画面400上で、キーボード503、マウス504などの入力デバイスを用いて様々な操作を行い、収集した文書を格納するフォルダを作成・消去したり、文書を収集するための検索条件を記述したり、収集結果を見たり、評価したりする。

【0033】文書収集クライアントシステム501が行う処理の流れを図7に示す。文書収集クライアントシステム501が起動されると、まず文書収集サーバシステムへの接続を行う（ステップ120）。次に図3に示すインタフェース画面400を表示する（ステップ121）。

【0034】この後、イベントループ122に入り、ステップ123～126を繰り返す。即ち、ユーザの操作を受け・解析し（ステップ123）、操作に対応するコマンドを文書収集サーバシステム100に送信し（ステップ124）、実行結果を文書収集サーバシステム100から受信し（ステップ125）、その実行結果をインタフェース画面400に反映させる（ステップ126）、という処理を繰り返す。

【0035】ユーザがメニューから終了を選ぶ操作を行うと、終了コマンドを文書収集クライアントシステム501に送信して、イベントループ122を抜け、文書収集サーバシステム100との接続切断処理を行い（ステップ127）、終了する。

【0036】図3に示したインターフェース画面400の具体例について説明する。この画面は、既にユーザによってフォルダ作成とそのフォルダに収集すべき文書の検索条件登録が行われ、文書収集サーバシステム100により、ユーザが登録した検索条件群に適合する文書を収集・分類された時点の例である。

【0037】402は、内容を表示中のフォルダの名前である。この例ではuser1というフォルダの下位ディレクトリであるvoiceというフォルダの内容を表示中である。

【0038】403にはフォルダuser1/voiceにユーザが登録した検索条件を表示する。表示されたテキストを直接編集することにより、検索条件の更新を行うことができる。本実施例では、各フォルダに対して記憶された検索条件は、単語（以下ワードと呼ぶ）、あるいは句、あるいは文章など、ユーザが自然語で記述し得るものを列挙したものからなる。

11

【0039】404にはフォルダuser1/voiceの下位のフォルダの一覧を表示する。各フォルダについて、フォルダ名、フォルダに収集されている文書数、フォルダに対応する検索条件の書き出しを表示している。この例では、user1/voiceの下にそれぞれ、recognitionとsynthesisの二つの下位フォルダがある。

【0040】この下位フォルダ一覧の項目をクリックするとクリックされた下位フォルダへ移動することができる。

【0041】405にはフォルダuser1/voiceにすでに収集されている文書の一覧を表示する。

【0042】各文書について、タイトル、フォルダuser1/voiceの検索条件への適合度、適合した検索条件中のワード、情報源名、ドメイン名などを表示している。

【0043】この文書一覧の項目をクリックすると、クリックされた文書の内容を見ることができる。文書の内容は406に表示される。

【0044】フォルダの作成・消去はメニュー401のFileメニューを使って行う。また、Goメニューを使っても、別のフォルダへ移動できる。

【0045】また、ユーザはメニュー401のEditメニューを使って収集された文書や文書が格納されているフォルダに対して評価を与えることができる。つまり、ユーザが、メニューを用いて、有用/無用な文書である、有用/無用なフォルダである、という評価を与えると、対応するコマンドが文書収集サーバシステム100に送られる。文書収集サーバシステム100は、文書やフォルダに対する評価を検索条件に反映させ、次の文書収集時からよりユーザの意図にあった文書を収集する。

【0046】サーバ510のメモリ上の文書収集サーバシステム100は、クライアント500からの要求を処理する一方で、ユーザが作成したフォルダ群と各フォルダに登録した検索条件に基づいて、文書の収集・分類・整理を行う。

【0047】つまり、文書収集サーバシステム100は、ニュースサーバ520や文書収集サーバ521などの外部の情報源に定期的にアクセスし、前回にアクセスした後で各情報源に蓄積された文書を取得し、ユーザが登録された検索条件群に適合するものを検索する。この際、検索条件中の各ワードの対象文書における出現数を対象文書とその検索条件との適合度とする。適合した検索条件に登録されているフォルダの中から、フォルダの階層構造を考慮して対象文書を分類するフォルダを選び、そのフォルダへ格納する。さらに、多くの文書が蓄積されたフォルダを自動分割するなどの文書の収集状況に応じた文書の整理を行う。

【0048】なお、文書収集の対象となる外部の情報源は、サーバからアクセス可能な他のネットワーク上に在っても良い。

12

【0049】文書の収集・分類・整理についてさらに詳しく説明する前に、まず、ユーザが作成するフォルダと検索条件について図4に示した例で説明する。

【0050】文書収集サーバシステム100にユーザ登録を行うと、各ユーザに一つのフォルダが割り当てられる。ユーザは自分に割り当てられたフォルダの下に、自由に、下位フォルダを階層的に作成して、各々のフォルダに対して、そのフォルダにはどのような文書を収集すべきかという検索条件を登録する。

【0051】図4の例では2人のユーザuser1、user2が登録されており、それぞれフォルダ540、フォルダ550が割り当てられている。user1は、フォルダ540の下に階層的にフォルダ541-544を作成し、各フォルダに検索条件545-548を登録している。

【0052】一方、user2は下位フォルダを作成せず、フォルダ550に、興味のある事柄を羅列しただけの検索条件551を登録している。

【0053】フォルダとフォルダに対応する検索条件は、ユーザが作成、更新するほか、文書収集サーバシステム100によっても、文書の収集状況に応じて自動的に作成されたり、更新されたりすることもある。詳しくは後述する。

【0054】したがって、user2のように、階層的なフォルダを作成せずに、興味のある事柄を羅列しておくだけでも、収集された文書は自動的に分類・整理される。

【0055】図2の流れ図に従い、文書収集サーバシステム100について説明する。

【0056】文書収集サーバシステム100は、複数のユーザからの要求にいつでも対応し、同時に定期的に文書の収集を行うために、常にユーザの接続要求があるか、または、一定時間が経過したかを監視している(ステップ101)。ユーザが接続要求をした場合には、クライアント要求処理110を開始する。一定時間が経過した場合には、文書収集処理102を開始する。いずれの場合も、文書収集サーバシステム100本体の処理は直ちにステップ101に戻り、ユーザの接続要求と一定時間経過の監視を続ける。

【0057】図6にクライアント要求処理110の流れとコマンド実行時に用いるデータ構造との対応を示す。

【0058】クライアント要求処理110が開始されるとまず、クライアントからの要求処理を行うための準備として、クライアントとの接続(ステップ111)、フォルダテーブルのロード(ステップ112)を行う。

【0059】このあと、クライアントから終了コマンドを受信するまで、クライアント500から送信されてくるコマンドの解析(ステップ113)と実行(ステップ114)を繰り返す。

【0060】終了コマンドを受信して、繰り返しを終了すると、クライアントの切断を行って、クライアント要求処理110を終了する。

13

【0061】各コマンドの実行時には、必要に応じて各種のテーブルのロード、参照、更新、セーブを行う。

【0062】たとえば、ユーザがあるフォルダに格納されている文書一覧の表示を要求する操作をすると、文書収集クライアントシステム501は対応するコマンドと対象のフォルダ名を送信する。クライアント要求処理110はこのコマンドとフォルダ名を受信すると、フォルダテーブルを参照して、そのフォルダに格納されている文書群の情報（各文書のタイトル、適合度、適合した検索条件中のワード、情報源名など）をクライアントへ送信する。

【0063】図5に示した文書収集処理102（図2）の流れと文書収集処理時に用いるデータ構造との対応に従って、文書収集処理について説明する。

【0064】まず、内部DB511からメモリ上に文書収集用のテーブル（文書番号テーブル300、フォルダテーブル310、ワード・フォルダテーブル330、ワード・文書テーブル350）をロードする（ステップ103）。

【0065】文書番号テーブル300は、どのような情報源が利用可能か、各情報源にはどのようなドメインがあるか、それらのドメインにはそれぞれ何番から何番までの文書があり、既に何番までは取得済みであるかという情報を表す。

【0066】フォルダテーブル310は、どのようなフォルダがどのような階層構造を成しているか、各フォルダにはどのような文書が格納されているかを表す。

【0067】ワード・フォルダテーブル330は、各フォルダに対応付けられている検索条件にはどのようなワードが出現するかを表す。

【0068】ワード・文書テーブル350にはどの文書にどのようなワードが出現するかを表している。各テーブルについて詳しくは後述する。

【0069】次に、各情報源の全ての新文書について、ステップ105～107を繰り返し実行する。

【0070】ステップ105の新文書取得処理は、各情報源に接続し、文書番号テーブル300に登録されている文書番号より新しい文書があるかどうか調べ、もしあればその文書を取得する。

【0071】次に、ステップ106の適合度計算が、取得した文書の各フォルダにおける適合度を計算する。まず、取得した文書にどのようなワードが出現するかを表すフォルダ検索テーブル370を作成し、各フォルダにおける適合度を記憶するために適合フォルダテーブル390を作成・初期化する。そして、フォルダ検索テーブル370とワード・フォルダテーブル330とを照合して、適合度をフォルダごとに算出し、適合フォルダテーブル390に登録する。適合度計算について詳しくは後述する。

【0072】次にステップ107の文書格納処理が、適

14

合フォルダテーブル390に登録された各フォルダにおける適合度と、フォルダテーブル310が表すフォルダ間の階層構造とから文書を格納するフォルダを決定し、その文書をフォルダテーブル310とワード・文書テーブル350に登録する。文書格納処理について詳しくは後述する。

【0073】次に、ステップ108のフォルダ管理処理が、ワード・文書テーブル350が表す各文書におけるワードの出現頻度分布を用いてフォルダ内の文書を分析し、フォルダの自動分割や統合を行ない、フォルダテーブル310とワード・フォルダテーブル330とを更新する。詳しくは後述する。

【0074】以上のステップ105～107の繰り返し中に更新された文書収集用テーブルを内部DB511へセーブする（ステップ109）。

【0075】ここまでで、一通りの文書収集処理102を終了する。

【0076】以上述べた文書収集処理102で用いるデータ構造や処理についてさらに詳しく説明する。

【0077】文書番号テーブル300のデータ構造を図13に示す。文書番号テーブル300は、ハッシュテーブルで、各エントリは図12に示す文書番号リスト302を指している。情報源名とドメイン名を入力とするハッシュ関数の値でエントリを決定する。

【0078】文書番号リスト302は、情報源名へのポインタ303、ドメイン名へのポインタ304、そのドメインの最古文書の番号305、最新文書の番号306、文書収集システムが既に収集処理を施した文書の番号307、同ハッシュ値の他の文書番号リストへのポインタ308の組である。

【0079】文書番号テーブルは、文書収集を始める際にロードされ、文書を情報源から取得する度に更新される。

【0080】内部DBには、どのような情報源があるか、どのようなドメインがあるか、どのドメインの文書は何番まで収集処理済みかが記憶されている。まず、内部DB511から、記憶されている情報源名、ドメイン名、既取得文書番号を読み込んで文書番号リスト302を作成し、情報源名とドメイン名を入力とするハッシュ関数の値をエントリとして文書番号テーブル300に登録する。次に、各情報源から各ドメインの最古文書番号、最新文書番号を取得し、文書番号リストに書き込む。このとき文書番号テーブル300に登録されていないドメインがあれば、これはその情報源において新規に作成されたドメインであるので、既取得文書番号を0として文書番号リストを生成し、文書番号テーブル300に登録する。

【0081】たとえば、図13の文書番号リスト302-aは、Internet news という情報源の fjal というドメインには、123番から145番までの文書があ

15

り、そのうち130番までが収集処理済みであることを示している。

【0082】フォルダテーブル310のデータ構造を図15に示す。フォルダテーブル310はハッシュテーブルで、各エントリは図14に示すフォルダリスト314を指している。フォルダ名を入力とするハッシュ関数の値でエントリを決定する。

【0083】フォルダリスト314は、フォルダのID番号315、フォルダ名へのポインタ316、上位フォルダを表すフォルダリストへのポインタ317、下位フォルダリスト321へのポインタ318、格納文書リスト324へのポインタ319、同ハッシュ値の他のフォルダを表すフォルダリストへのポインタ320の組である。

【0084】下位フォルダリスト321は、下位フォルダを表すフォルダリストへのポインタ322とフォルダリスト314で表されるフォルダの他の下位フォルダを表す下位フォルダリストへのポインタ323の組である。

【0085】格納文書リスト324は、格納された文書の情報源名へのポインタ325、ドメイン名へのポインタ326、文書番号327、格納文書リスト324が表す文書のフォルダリスト314が表すフォルダにおける適合度328、このフォルダに格納された他の文書を表す格納文書リストへのポインタ329の組である。

【0086】例えば、図15のフォルダリスト314-aは、フォルダIDが1003のvoiceというフォルダの上位フォルダはフォルダリスト314-bで表されるフォルダuser1であること、フォルダリスト314-cで表されるフォルダsynthes1を下位フォルダに持つこと、このフォルダには適合度13点のinternet newsという情報源のfj.aiというドメインの120番の文書等が格納されていることを表している。

【0087】図17に示すワード・フォルダテーブル330は、ハッシュテーブルで、各エントリは図16に示すワード・フォルダリスト333を指している。ワードを入力とするハッシュ関数の値でエントリを定める。

【0088】ワード・フォルダリスト333は、ワードへのポインタ334、フォルダ頻度リスト340へのポインタ335、同ハッシュ値の他のワード・フォルダリストへのポインタ336の組である。フォルダ頻度リスト340は、このワードが出現する検索条件に対応するフォルダのフォルダID341、検索条件中のワードの出現頻度342、他のフォルダ頻度リストへのポインタ343の組である。

【0089】例えば、図17のフォルダリスト333-aとフォルダ頻度リスト340-aは、言語というワードが、フォルダID1003のフォルダに対応する検索条件中に1回出現することを表し、フォルダリスト333-bとフォルダ頻度リスト340-b、340-c

16

は、音声認識というワードが、ID1003のフォルダとID1004のフォルダのそれぞれに対応する検索条件中に1回ずつ出現することを表す。

【0090】ワード・文書テーブル350のデータ構造を図19に示す。ワード・文書テーブル350はハッシュテーブルで、各エントリは図18に示すワード・文書リスト354を指している。ワードを入力とするハッシュ関数の値でエントリを決定する。

【0091】図18のワード・文書リスト354は、ワードへのポインタ355、文書頻度リスト360へのポインタ356、同ハッシュ値の他のワード・文書リストへのポインタ357の組である。文書頻度リスト360は、このワードが出現する文書の情報源名へのポインタ361、ドメイン名へのポインタ362、文書番号363、出現頻度364、このワードが出現する他の文書頻度リストへのポインタ365の組である。

【0092】例えば、図19のワード・文書リスト34-aと文書頻度リスト360-aは、言語というワードが、情報源internet newsのドメインfj.sci.langの56番の文書に5回出現することを表し、ワード・文書リスト34-bと文書頻度リスト360-b、360-cは、音声認識というワードが情報源internet newsのドメインfj.aiの120番の文書に2回出現し、ドメインfj.sci.langの56番の文書に2回出現することを表している。

【0093】フォルダテーブル310、ワード・フォルダテーブル330、ワード・文書テーブル350の内容は、内部DB511に記憶されている。これらのテーブルは文書収集処理102が開始されたときクライアント要求処理110が開始された時やコマンド実行時に、必要に応じてメモリ上へロードされ、それぞれの処理を実行中に参照・更新され、終了するときに内部DB511にセーブされる。ただし、各テーブルは排他的に更新される。フォルダの作成・削除によるフォルダテーブルの更新、検索条件の更新によるワード・フォルダテーブルの更新は、ただちにセーブされる。

【0094】例としてワード・フォルダテーブル350のロードについて図8に流れ図を示す。ワード・フォルダテーブル350のロードは、フォルダテーブル310をロードした後で行う。

【0095】まず、ワード・フォルダテーブル350を初期化する(ステップ160)。

【0096】次に、フォルダテーブル310に登録されている全てのフォルダについて、フォルダのワード登録(ステップ164~166)を繰り返す(ステップ161)。

【0097】フォルダのワード登録は、まず、そのフォルダに対応する検索条件を内部DB511からメモリ523上に読みこみ、(ステップ164)、ワードを抽出する(ステップ165)。抽出した各ワードについて図

17

16のワード・フォルダリスト333を作成し、ワードのハッシュ値を計算して図17のワード・フォルダテーブル330に登録する(ステップ166)。

【0098】全てのフォルダについてワード登録を行うとの繰り返しを終了し、ワード・フォルダテーブルロード処理151を終了する。

【0099】適合度計算106が行なう検索処理について図9に基づいて説明する。

【0100】この処理は、検索条件群に出現するワードと文書に出現するワードの類似性を調べることに、

取得した文書と各フォルダの適合度を調べる。

【0101】ここで本実施例で使用する、検索対象文書といずれかのフォルダとの適合度について説明する。

【0102】検索対象文書といずれかのフォルダとの適合度は、いくつかの方法が考えられるが、本実施例では、その文書内のワードのうち、そのフォルダに適合したワード(すなわち、そのフォルダに対応して記憶された検索条件に含まれるワードに一致した、文書内のワード)のそれぞれとそのフォルダとの適合度を求め、それらのワードとそのフォルダとの適合度の総和を求め、この総和をその文書とそのフォルダとの適応度とする。

【0103】ここで、そのフォルダに適応したワードとそのフォルダとの適応度もいろいろの方法で求めることが出来るが、本実施例では、より好適なものとして、そのワードとその文書内の重みとそのワードのそのフォルダ内での重みとの積をもってそのワードとそのフォルダの適応度とする。

【0104】ここで、そのワードの文書内の重みは、いろいろの方法で検出可能であるが、本実施例では、より好適なものとして、そのワードのその文書内での出現頻度をもって、そのワードのその文書内での重みとする。

【0105】さらに、そのワードとそのフォルダとの適応度もいろいろの方法で検出可能であるが、本実施例では、より好適なものとして、そのフォルダに対応して記憶された検索条件内でのそのワードの出現回数を使用する。

【0106】従って、本実施例では、そのワードとそのフォルダとの適合度は、そのワードの文書内出現頻度とそのワードのそのフォルダに対応する検索条件内での出現頻度の積をもって表すことが出来、その検索対象文書とそのフォルダとの適応度は、このようにして求めた各ワードの適応度の総和で与えられる。

【0107】より具体的には、取得した文書に出現するワードを図21のフォルダ検索テーブル370に登録し、全フォルダに対応する検索条件に出現するワードを登録したワード・フォルダテーブル330と照合して、フォルダ毎に適合度を集計し、適合度順にフォルダをソートする。

【0108】まず、フォルダ検索テーブル370の初期化(ステップ170)、図23の適合フォルダテーブル

18

390の初期化(ステップ171)を行なう。

【0109】次に取得文書からワードを抽出し(ステップ172)、各ワードをフォルダ検索テーブル370に登録する(ステップ173)。

【0110】フォルダ検索テーブル370はハッシュテーブルで、各エントリは、図20に示すフォルダ検索リスト372を指す。ワードを引数とするハッシュ関数の値でエントリを決定する。フォルダ検索リスト372は文書中のワードへのポインタ373、適合フォルダリスト380へのポインタ374、文書中の出現頻度375、同ハッシュ値の他のフォルダ検索リストへのポインタ376の組である。適合フォルダリスト380は、ワードが出現する検索条件に対応するフォルダのフォルダID 381、そのフォルダにおける適合度382、他の適合フォルダリストへのポインタ383の組である。

【0111】たとえば、図21のフォルダ検索テーブル330のフォルダ検索リスト372-aは、言語というワードが検索対象の文書中に2回出現することを表している。まだ検索を実行していないので、適合フォルダリストへのポインタ374-aはNULLである。同様に、フォルダ検索リスト372-b、372-cはそれぞれ対象文書中に無音時間というワードが3回出現すること、音声認識というワードが5回出現することを表している。

【0112】次に、フォルダ検索テーブル370とワード・フォルダテーブル330を照合し、適合するフォルダがあれば、適合フォルダリストを作成し、フォルダ検索テーブルに登録する(ステップ174)。

【0113】すなわち、フォルダ検索テーブル370に登録されているワードが、ワード・フォルダテーブル330にも登録されていれば、フォルダ検索リスト372の頻度375とワード・フォルダリスト333に登録されている各フォルダ頻度リスト340の頻度342を掛け合わせた値をもってそのワードの各フォルダにおける適合度として、それぞれに対応する適合フォルダリスト380を作成し、フォルダ検索リスト372に登録する。

【0114】例えば、フォルダ検索テーブル370に登録されている言語というワードはワード・フォルダテーブル330のフォルダ頻度リスト340-aが示すように、フォルダID 1003に対応する検索条件に1回出現している。したがって、言語 というワードのID 1003のフォルダにおける適合度は2点で、検索実行前にはNULLであった適合フォルダリストへのポインタは、図22に示した検索実行後のフォルダ検索テーブル370のように適合フォルダリスト380-aを指す。

【0115】同様に音声認識というワードは、ワード・フォルダテーブル330のフォルダ頻度リスト340-bが示すように、ID 1003のフォルダとID 1004のフォルダに対応する検索条件にそれぞれ1回出現している。したがって、音声認識というワードのID 1003

19

のフォルダとID1004のフォルダにおける適合度はそれぞれ5点である。したがって、検索実行後は、図22のフォルダ検索テーブル370に適合フォルダリスト380-bと380-cが登録される。しかし、無音時間というワードは、ワード・フォルダテーブルに登録されていない。すなわち適合するフォルダが存在しないということで、検索実行後も適合フォルダリストへのポインタはNULLである。

【0116】最後にフォルダ毎に、各ワードのフォルダにおける適合度を集計し、適合度が0でないフォルダを適合度の高い順に図23の適合フォルダテーブル380に登録する(ステップ175)。適合フォルダテーブルの各エントリは、適合フォルダリスト381を指す。適合フォルダリストは、図20の適合フォルダリストと同じデータ構造であるが、各ワードのフォルダにおける適合度の合計を適合度とし、他の適合フォルダリストへのポインタ383は使用しない。

【0117】以上の適合度計算によって作成された適合フォルダテーブル390とフォルダの階層構造が登録されているフォルダテーブル310を用いて、文書格納処理109が、文書を格納すべきフォルダを選んで格納する。文書格納処理109の流れを図10に示す。

【0118】この文書格納処理は、大きく分けて2段階(ステップ180、181)からなる。

【0119】まず、ステップ180では、適合フォルダテーブル390に登録された各フォルダにおける適合度とフォルダテーブル310に登録されたフォルダの階層構造から、対象文書をどのフォルダに格納すべきかを決定し、対象文書を格納した文書としてフォルダテーブル310に登録する。

【0120】図24に格納フォルダの決定方法の説明図を示す。本実施例ではフォルダの階層構造の各枝で適合したフォルダの中で最も下位のフォルダに格納する。AからHまでのフォルダがあり、それぞれ図に示す適合度であった場合には、A-B-D-Gという枝の適合したフォルダの中で最も下位のD、同様に枝A-B-E中のE、枝A-C-E-H中のEに格納する。この方法は、下位のフォルダは上位のフォルダの検索条件を継承していると考え、検索条件をより詳しく記述しているフォルダに格納するものである。

【0121】次に、ステップ181では、対象文書に出現するワードをワード・文書テーブル350に登録する。

【0122】すなわち、文書に出現する全ワードについて図18のワード・文書リストを作成し、ワード・文書テーブル350に登録する。

【0123】この処理により、ワード・文書テーブル350には格納された全文書について、各文書にどのようなワードが出現するかが記録される。このワード・文書テーブル350は、次に述べるフォルダ管理処理でフォ

20

ルダ内の文書を分析するのに用いる。

【0124】フォルダ管理処理108について説明する。

【0125】階層構造を成すフォルダに対応する検索条件を分類体系とみなして文書の収集、分類を続ける。文書が特定のフォルダに集中して、フォルダ内の文書数がユーザが把握しきれないほど増えることがある。また、文書が複数のフォルダに重複して格納されることが多くなり、無駄が生じることもある。

【0126】これらの現象は、ユーザが適切に分類体系を構成していなかった場合や世間の情勢や研究動向が変化し、分類体系が合わなくなった場合に起きる。

【0127】フォルダ管理処理108は、各フォルダへの文書の集まり具合を分析することによって、これらの現象を検知し、フォルダの階層構造やフォルダに対応する検索条件を改良する。これにより、フォルダ内の文書数がユーザが把握できる程度に抑ええたり、文書が複数のフォルダに不必要に重複して格納されないようにしたりし、フォルダの階層構造を実情にあった体系に維持する。

【0128】フォルダ管理処理の流れを図11に示す。

【0129】各フォルダに対してステップ201へステップ205を繰り返し行なう(ステップ200)。

【0130】まず、フォルダに格納された文書数を監視する(ステップ201)。フォルダにあらかじめ与えた数以上の文書が格納されていれば、そのフォルダ内の文書を統計的手法を用いて分析する(ステップ202)。

異なった性質のものが混ざり合っている対象の中で、類似している個体を集めてグループに分類する手法はクラス分析として知られており、たとえば、『多変量解析ハンドブック』(現代数学社1986年)に記載されている。ステップ202はクラス分析の手法を用いて、フォルダ内の文書に出現するワードの頻度に基づき文書を再分類する。再分類した文書の集合をクラスと呼ぶ。

【0131】次にクラス間関係を分析し、クラス間の階層構造を決定する(ステップ203)。ここで行なうクラス間関係解析については後述する。クラスタに対応してフォルダと検索条件を生成し、クラス間の階層構造に対応して階層的にフォルダを作成する(ステップ204)。次にワード・文書テーブル350から、生成した各フォルダ内の文書に共通して高頻度に出現するワードを抽出し、フォルダに対応する検索条件に加える(ステップ205)。これにより、検索条件を精練することができる。

【0132】ここまでの処理を各フォルダに施したら、各フォルダに格納された文書群を分析し、フォルダの再構成、すなわち、フォルダの統合、階層構造の変更を行なう(ステップ206)。ここで行なう分析については後述する。

21

【0133】図26-31を使ってステップ206で行なうクラスタ間関係の分析方法を説明する。

【0134】図25のように、ワード群w1とワード群w2からなる検索条件があり、この検索条件に対応するフォルダ450と文書群dが格納されているとする。このとき、このフォルダ内の文書について、ワード・文書テーブルから得られるデータを統計的に分析してえらる、ワードと文書の関係のパターンを図26の451、図26の455、図28の458に示す。

【0135】図26は、文書群dがワード群w1が出現する文書群d1とワード群w2が出現する文書群d2の二つの独立したクラスタに分類される場合である。この場合、ワード群w1からなる検索条件とワード群w2からなる検索条件を生成し、それぞれに対応するフォルダ453、454と両者の上位のフォルダ452を設け、図26に示す階層構造にする。

【0136】図27は、ワード群1のみが出現する文書群d1とワード群w1とワード群w2が出現する文書群d2の二つのクラスタに分類される場合である。ワード群w2が出現する文書群にはワード群w1も出現している。そこで、ワード群w1からなる検索条件とワード群w2からなる検索条件を生成し、それぞれに対応するフォルダ456と457を設け、図27に示す階層構造にする。

【0137】図28は、ワード群w1のみが出現する文書群d1とワード群w2のみが出現する文書群d3とワード群w1とワード群w2の両方が出現する文書群d2の3つのクラスタに分類される場合である。この場合、ワード群w1のみからなる検索条件とワード群w2のみからなる検索条件とワード群w1かつワード群w2なる検索条件を生成し、それぞれに対応するフォルダ456、457、458とこれらの上位のフォルダ455を設け、図28のような階層構造にする。

【0138】同じ図25-30を使ってステップ205で行なうフォルダ間関係の分析方法を説明する。

【0139】図27のような階層構造のフォルダがあるときに、フォルダ453とフォルダ454に重複して格納される文書が増えたとなると、ワードと文書の関係のパターンに変化しと考えられる。文書が重複して格納されていることは、フォルダテーブル310から検知できる。そこで、フォルダ453、454に格納されている文書について、ワード・文書テーブル350から得られる各文書におけるワードの出現頻度に基づきクラスタ間関係分析と同様の統計的分析を行ってワードと文書の分布のパターンを調べ、パターンに応じてフォルダと検索条件を再構成する。

【0140】また、図27のような階層構造があるときに、フォルダ457に格納される文書の中でフォルダ456に適合しない文書の割合があらかじめ与えられた割

22

合を越えるようになった場合、検索条件の上下関係が変化したことを意味する。このことは、フォルダテーブル310に登録されているフォルダ457に格納された文書のフォルダ456への適合度を調べることににより検知できる。フォルダ456とフォルダ457とに格納されている文書に対しクラスタ分析を行ない、フォルダを再構成する。

【0141】

【発明の効果】本発明によれば以下の効果が得られる。

【0142】(1) ユーザが記述した検索条件群に適合する情報を複数の情報源から収集し、検索条件群の階層構造を分類体系と見做して収集した情報を分類できる。

【0143】(2) 各検索条件に対応した検索結果格納領域への情報の集まり具合に応じて、分類体系を変更することができる。

【0144】その結果、ある検索結果格納領域に格納される情報量をユーザがその全体を把握できる程度の数に抑えることができる。あるいは情勢の変化に応じ、適切な分類体系を維持できる。

【図面の簡単な説明】

【図1】本発明の一実施例のシステム構成図である。

【図2】本実施例の文書収集サーバシステムの流れ図である。

【図3】本実施例のインタフェース画面の例である。

【図4】本実施例のフォルダとフォルダに対応する検索条件の例である。

【図5】本実施例の文書収集処理の流れ図である。

【図6】本実施例のクライアント要求処理の流れ図である。

【図7】本実施例の文書収集クライアントシステムの流れ図である。

【図8】本実施例のワード・フォルダテーブルロード処理の流れ図である。

【図9】本実施例の適合度計算の流れ図である。

【図10】本実施例の文書格納処理の流れ図である。

【図11】本実施例のフォルダ管理処理の流れ図である。

【図12】本実施例の文書番号リストのデータ構造である。

【図13】本実施例の文書番号テーブルのデータ構造である。

【図14】本実施例のフォルダリストと下位フォルダリストと格納文書リストのデータ構造である。

【図15】本実施例のフォルダテーブルのデータ構造である。

【図16】本実施例のワード・フォルダリストとフォルダ頻度リストのデータ構造である。

【図17】本実施例のワード・フォルダテーブルのデータ構造である。

【図18】本実施例のワード・文書リストと文書頻度リ

23

ストのデータ構造である。

【図19】本実施例のワード・文書テーブルのデータ構造である。

【図20】本実施例のフォルダ検索リストと適合フォルダリストのデータ構造である。

【図21】本実施例のフォルダ検索テーブルのデータ構造である。

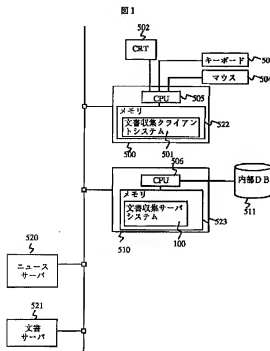
【図22】本実施例の検索処理後のフォルダ検索テーブルの例である。

【図23】本実施例の適合フォルダテーブルのデータ構造である。

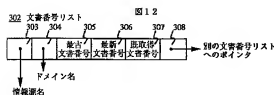
【図24】本実施例の文書を格納するフォルダの決定方法の説明図である。

【図25】本実施例のフォルダ内文書分析を行なうフォルダの説明図である。

【図1】



【図12】



24

【図26】本実施例のフォルダ内文書分析結果のワードと文書の第1の分布ボタンそれぞれに従って生成されるフォルダの説明図である。

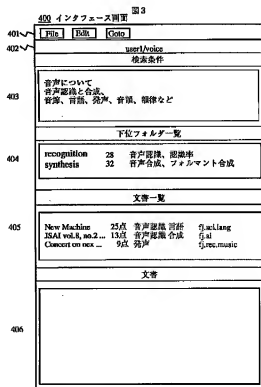
【図27】本実施例のフォルダ内文書分析結果のワードと文書の第2の分布ボタンそれぞれに従って生成されるフォルダの説明図である。

【図28】本実施例のフォルダ内文書分析結果のワードと文書の第3の分布ボタンそれぞれに従って生成されるフォルダの説明図である。

【符号の説明】

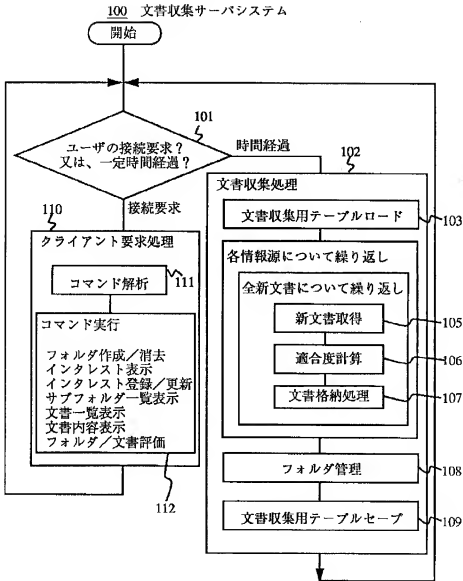
100…文書収集サーバシステム、102…文書収集処理、106…適合度計算、107…文書格納処理、108…フォルダ管理処理、110…クライアント要求処理。

【図3】

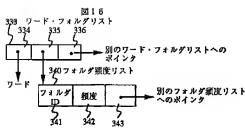


【図2】

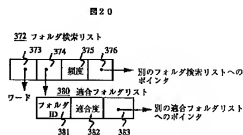
図2



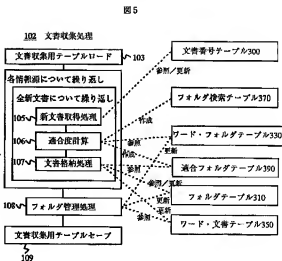
【図16】



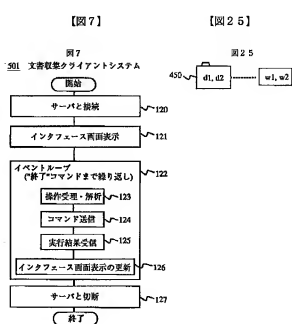
【図20】



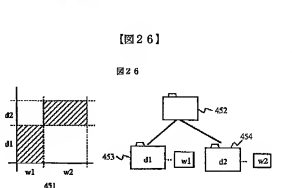
【図5】



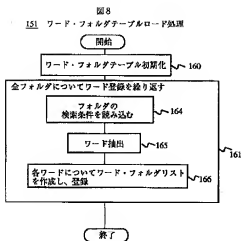
【图 2 5】



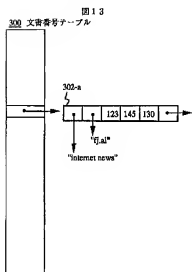
【图 2-6】



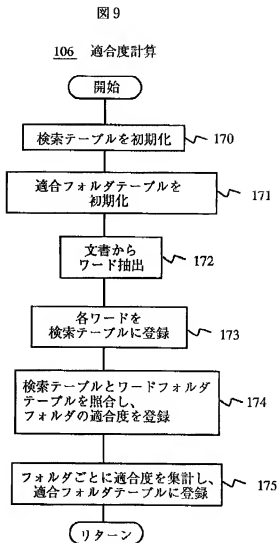
【図8】



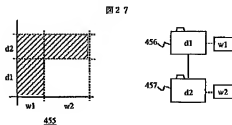
【図13】



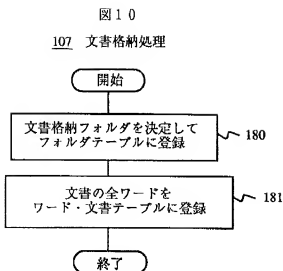
【図9】



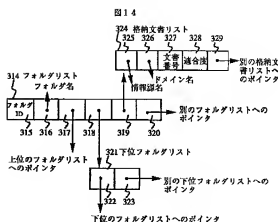
【図27】



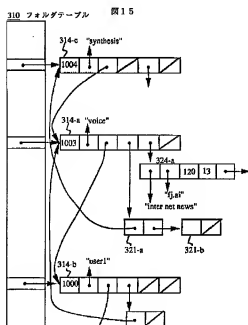
【図10】



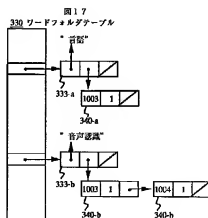
【図14】



【図15】



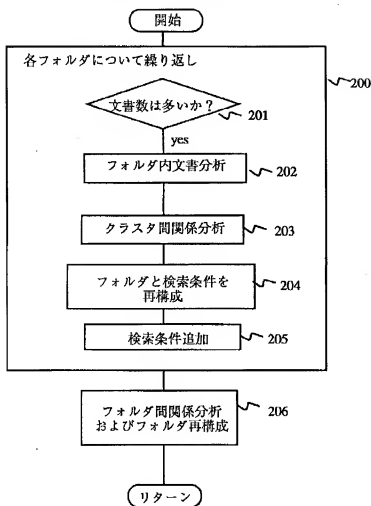
【図17】



【図11】

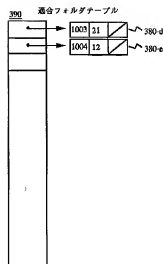
図11

108 フォルダ管理処理



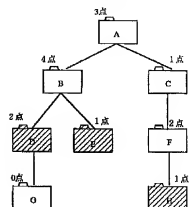
【図23】

図23

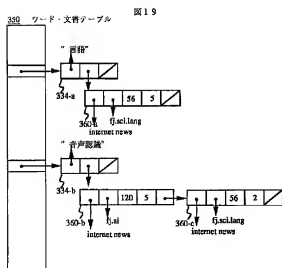


【図24】

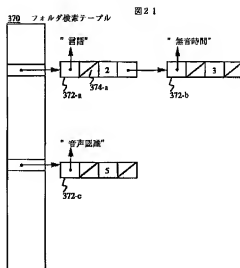
図24



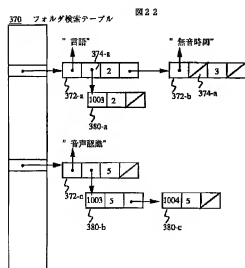
【図19】



【図21】



【図22】



【図28】

